用例翻訳ＳＤＭＴにおける用例数の効果

加藤　直人　　渦原　茂
ＡＴＲ音声言語コミュニケーション研究所
{naoto.kato, shigeru.uzuhara}@atr.jp

1　はじめに

用例翻訳ＳＤＭＴ（Similarity-Driven Machine Translation）の研究を行っている[1][2]。ＳＤＭＴは、単言語内類似度と二言語間類似度を柱とした翻訳方式であり、従来の機械翻訳のような言語変換処理を持たないという特徴がある。また、翻訳には対訳用例を必要とするが、統計翻訳のように事前に翻訳知識を学習する必要がない。

ＳＤＭＴでは用例を組合せることで翻訳を行うので、翻訳精度を向上させる単純な方法は用例数を増やすことである。本稿では、旅行対話を対象として、ＳＤＭＴにおいて用例数の増加が精度に与える効果について報告する。

2　ＳＤＭＴ

まず、我々が提案するＳＤＭＴについて、その概要を説明する。以下では、旅行対話の日英翻訳を例に取って説明する。旅行対話では類似した表現がよく使われるので、用例翻訳が適していると考えられる。

本手法は大きく分けて、以下の４つのステップから構成されている。

STEP1:単言語内類似度による用例の収集
STEP2:マルチプルアライメント
STEP3:ワードグラフによる解候補の構成
STEP4:言語モデルと二言語間類似度による最適解の探索

STEP1は原言語側での処理であり、STEP2～STEP4は目的言語側での処理である。このように原言語側での処理と目的言語側での処理が分かれており、両言語にわたる言語変換処理がない。

STEP1は、入力文と類似した用例の収集を行う。これは用例翻訳一般が行う処理である。文間の類似度は、ＳＤＭＴでは単言語内類似度と呼び、後述するように目的言語側でも計算する。ここで、単言語内類似度は、２つの文S_1,S_2間に共通する単語数の割合、

$$sim(S_1,S_2) = 2 \cdot |S_1 \cap S_2| / (|S_1| + |S_2|) \quad (1)$$

（|・|は単語数を表す）

で定義している。ただし、共通単語数を求める際には、まず２文間で単語アライメントを行い、一致した単語を共通の単語としている。単言語内類似度は0から1.0までの値を取り、1.0に近いほど２つの文が類似している。

ＳＤＭＴで収集する用例は目的別に、**基本用例**（S_{base}）、**単語用例**（S_{words}）、**補足基本用例**（$S_{baseplus}$）、**補足被覆単語用例**（$S_{wordsplus}$）という４種類である。基本用例は出力文の骨格を構成するためのものである。これは、入力文に類似した文から上位$|S_{base}|$個を収集する。単語用例は基本用例で被覆できなかった単語を収集するためのものである。具体的には、基本用例で被覆できなかった単語を含み、かつ、単言語内類似度が高い上位$|S_{words}|$個を収集する。例えば、次の入力文を日英翻訳することを考えると、

【入力文】
J_0: グラスゴーまで寝台の切符をお願いします

翻訳用例として、図１のS_{base} = {用例１，用例２，用例３}（$|S_{base}|$ = 3）が、そして、基本用例で被覆されない単語“寝台”を補うために、S_{words} = {用例４}（$|S_{words}|$ = 1）が収集される。

さらに、残った用例から同様な方法で再度、基本用例と単語用例を収集したのが、補足基本用例と補足単語用例である。基本用例と単語用例はSTEP3のワードグラフの作成とともにSTEP4の二言語間類似度の計算に利用されるのに対して、補足基本用例と補足単語用例はSTEP4の二言語間類似度の計算のみで利用される。

【用例１】

J_1 ：サンフランシスコまでの片道切符をお願いします　E_1 ：I ‘d like a one-way ticket to San Francisco, please.

【用例2】

J_2 ：グラスゴーまで特急をお願いします　E_2 ：I would like a ticket on the limited express to Glasgow, please.

【用例3】

J_3 ：ロンドンまでの指定券をお願いします　E_3 ：I ‘d like to reserve a seat to London.

【用例4】

J_4 ：シカゴ行き寝台の切符を二枚ください　E_4 ：Can I have sleeping car tickets to Chicago, please?

図１　単言語内類似度によって収集された用例

STEP2では、収集された用例における目的言語側の文を、単語レベルで組み合わせる。その組み合わせ方はマルチプルアライメントで決定する。ここで、マルチプルアライメントとは配列を比較し並べる方法である。マルチプルアライメントは、ペアワイズアライメントを３本以上の文字列比較に拡張したものであり、機械翻訳に応用されている例もある[3]。マルチプルアライメントの具体的な手法にも様々あるが、今回は、グローバルアライメント、アフィンギャップを使い、ツリーベース組合せ法を利用した。図１の英語文をマルチプルアライメントすると図２のような結果が得られる。

STEP3では、解候補をワードグラフで構成する。まず、同一直線上にアライメントされた単語を１つのノードにまとめ、それ以外の単語は単独でノードにする。そして、隣接するノードをアークで結ぶ。図２のアライメント結果からは、図３のワードグラフが作成される。

STEP4では、ワードグラフから最適パスを探索することにより最適解（出力文）を得る。最適パスは言語モデルと二言語間類似度の制約を用いて求める。

言語モデルの制約では、目的言語でのn-gram（実際にはtrigram）を用いて、パスのNベスト集合（E_N）求める。

次に二言語間類似度の制約を用いて、言語モデルで得られたNベスト集合（E_N）を再ランク付けする。ここで、二言語間類似度は

$$\Delta_i = 1 - |sim(J_0, J_i) - sim(E_0, E_i)| \qquad (2)$$

と定義している。また、二言語間類似度に関して、次のような大胆な仮定をする。

〔仮定〕

２つの用例間において、原言語と目的言語の単言語内類似度はそれぞれほぼ同じ値となる。

E_1:		I	'd	like		a	one-way	ticket		to	San Francisco	,	please	.	
E_2:		I	would	like		a		ticket	on the express	to	Glasgow	,	please	.	
E_3:		I	'd	like	to reserve	a	seat			to	London			.	
E_4:	Can	I			have		sleeping	car	tickets	to	Chicago	,	please	?	

図２ 目的言語側の用例をマルチプルアライメントした結果

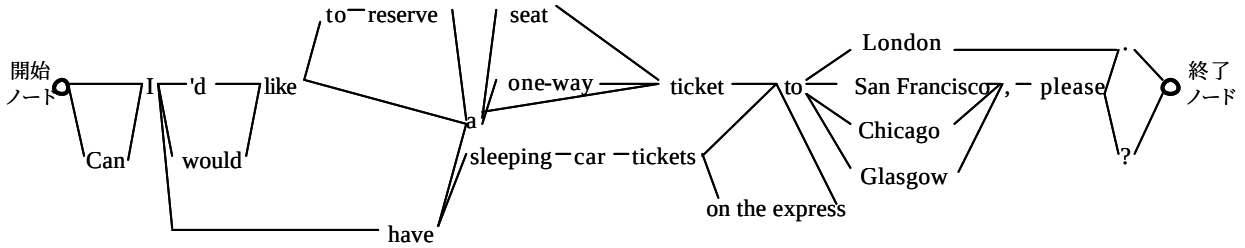

図３ 解候補を表すワードグラフ

この仮定を次のように定式化した。収集された用例をSとすると、最適解はS中の二言語間類似度の総和が最大となるものがよいと考えられるので、次式で定義することが考えられる。

$$\hat{E} = \underset{E \in E_N}{\operatorname{argmax}} \frac{1}{|S|} \sum_{i \in S} \Delta_i \qquad (3)$$

また、式(3)の精度を高めるためにはなるべく多くの用例Sを収集したほうがよい。しかしながら、多くの用例を収集すると、ワードグラフが大きくなってしまい、探索に時間がかかるという問題が生じる。そこで、式(4)のように、用例をワードグラフの作成で用いるもの（基本用例と単語用例）とそうでないもの（補足基本用例と補足単語用例）とに分けた。

$$\hat{E} = \underset{E \in E_N}{\operatorname{argmax}} \left(m_1 \frac{1}{|S_{base} \cup S_{words}|} \sum_{i \in S_{base} \cup S_{words}} \Delta_i + (1-m_1) \frac{1}{|S_{baseplus} \cup S_{wordsplus}|} \sum_{j \in S_{baseplus} \cup S_{wordsplus}} \Delta_j \right) \qquad (4)$$

これらの２つ制約により最適パスを探索すると、例えば次のような出力文が得られる。

【出力文】

E0: I’d like a sleeping car ticket to Glasgow, please.

３　実験

本手法の評価実験を用例数を変えて行った。テストセットはIWSLT2004[4]のテストセット500文を用いた。用例はIWSLT2004で与えられている2万文(日本語文は単語に分かち書きされている)の日英対訳を初期値とし、ATRで開発している旅行対話基本用例集BTEC[5]を加えて、５万分、７万文、10万文、12万文、15万分、17万文、20万文、‥、45万文と増やした。翻訳評価は自動評価指標であるBLEU，NIST，WER，PER，GTMで行った。

パラメータは次のようにした。単語間類似度は表層形が一致したときは1.0、不一致のときは0.0と定義した。収集する４種類の用例の数はそれぞれ、$|S_{base}| = 4$、$|S_{words}| = 1$、$|S_{baseplua}| = 1$、$|S_{wordsplus}| = 1$とした。２つの用例集合の重みμ_1は0.7とした。

図４～７に各評価指標ごとの結果を示す。

図４～７を見ると、BLEU,NIST,WER,PER,GTMのいずれもが用例数の増加とともに良くなる傾向にあることがわかる。また、用例数が30万文を越えるとその改善が鈍化している。精度向上の要因の１つに、入力文と用例数の類似度が高くなっていることが考えられる。そこで、入力文と用例数の平均類似度が用例数とともにどれくらい変化するかを調べた。その結果を図８に示す。図８を見ると、用例数の増加とともに平均類似度が高くなっていくことが

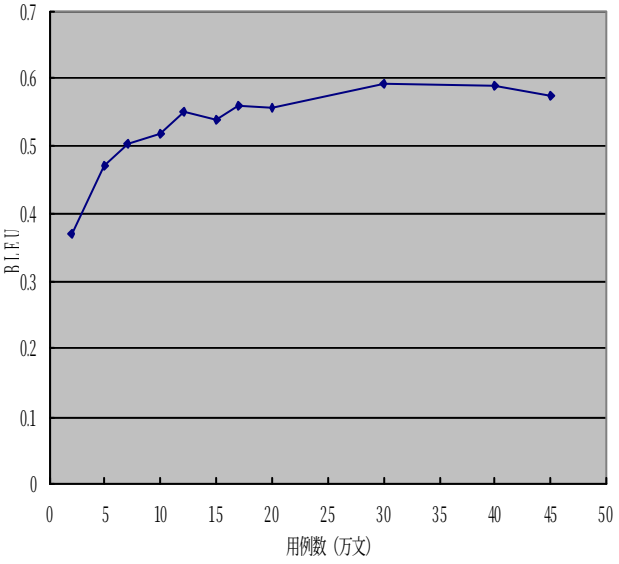

図４　BLEUと用例数の関係

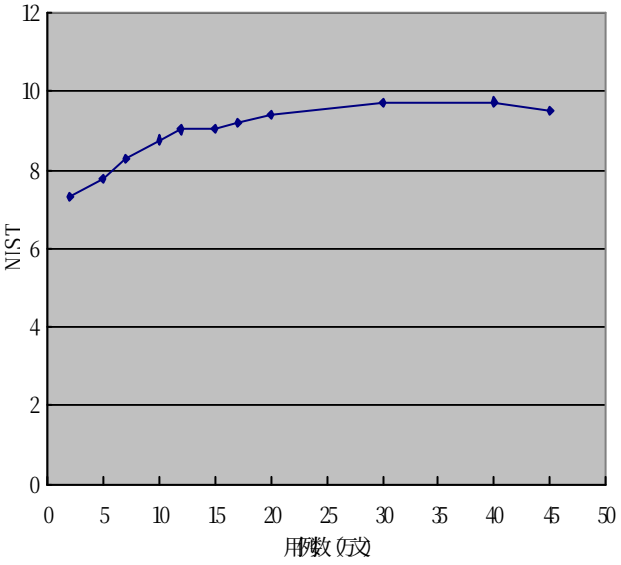

図４　NISTと用例数の関係

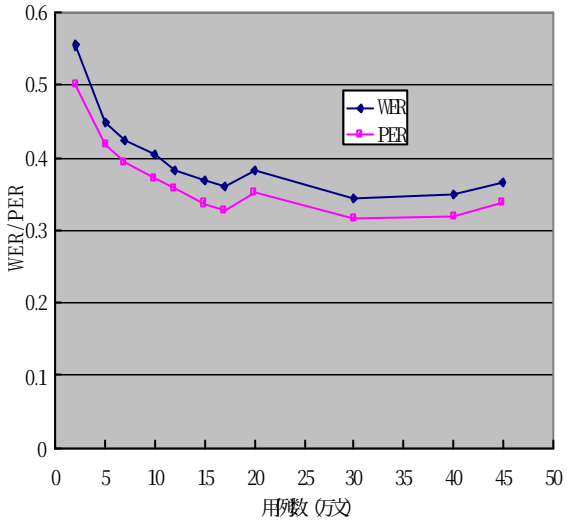

図６ WER/PERと用例数の関係

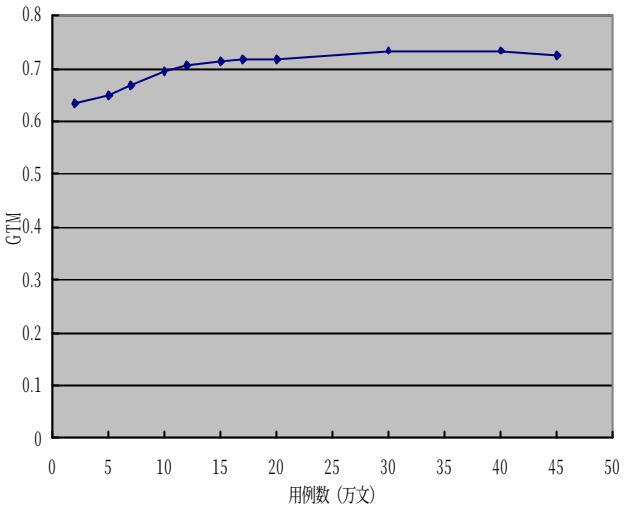

図７ GTMと用例数の関係

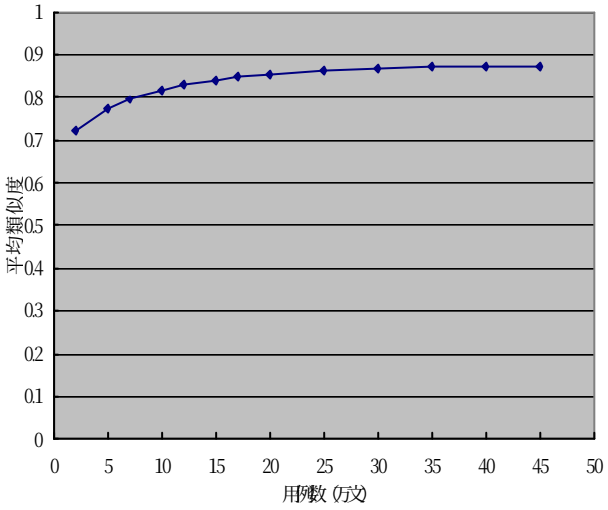

図８平均類似度と用例数の関係

わかる。また、平均類似度も用例数が30万文で飽和している。

４　おわりに

用例翻訳ＳＤＭＴと、その用例数を増やしたときの効果について述べた。今回は、ＳＤＭＴの改善方法として単純にコーパスを増やすことを考えたが、その他の言語資源を使うことも考えられる。例えば、目的言語で品詞やシソーラスを使うことにより、マルチプルアライメントの精度を向上させることができる。また対訳辞書を使うことにより、コーパスには出現しない単語を補ったり,原言語と目的言語間で単語のアライメントを付け、最適解の探索時にそのパスを優先することにより良い解が得られることが期待できる。

今後はこのような改善を行ってＳＤＭＴの精度を向上させ、IWSLTのUnristricted Data Trackでの比較を行う。

謝辞

本研究は独立行政法人 情報通信研究機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである。

参考文献

[1]加藤直人．マルチプルアライメントによ用例翻訳．第４回科学技術フォーラムFIT2005，pp.175-178，2005.
[2]加藤直人．ＳＤＭＴ：用例翻訳への新しいアプローチ．情報処理学会自然言語処理研究会，NL-170，pp.151-156，2005.
[3]Srinivas Bangalore, et.al. Bootstrapping Data using Consensus Translation for a Multi-lingual Instant Message System. Proc. of COLING02, 2002.
[4]Yasuhiro Akiba, et.al. Overview of the IWSLT04 Evaluation Campaign. Proc. of IWSLT04, pp.1-12, 2004.
[5]中岩浩巳．多言語翻訳技術に関する公開性能評価 ― 音声翻訳技術のための国際評価ワークショップIWSLT2004，電子情報通信学会第6回音声言語シンポジウム，SLP-54，pp.133-138，2004.